Newswire: A Large-Scale Structured Database of a Century of Historical News

Conference on Neural Information Processing Systems

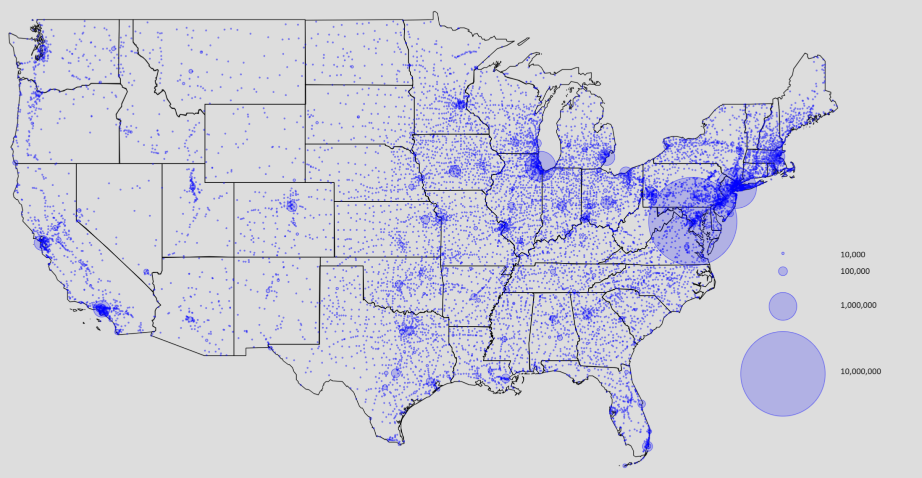
Emily Silcock, Abhishek Arora, Luca D’Amico-Wong, Melissa Dell

Dataset Overview

- 2.7 million unique public domain U.S. newswire articles
- Spans 1878-1977
- Each article reproduced 32 times on average
- Includes georeferencing, topic tagging, named entities, and individual disambiguation.
- Lists newspapers that printed the article on their front page, with Library of Congress metadata.

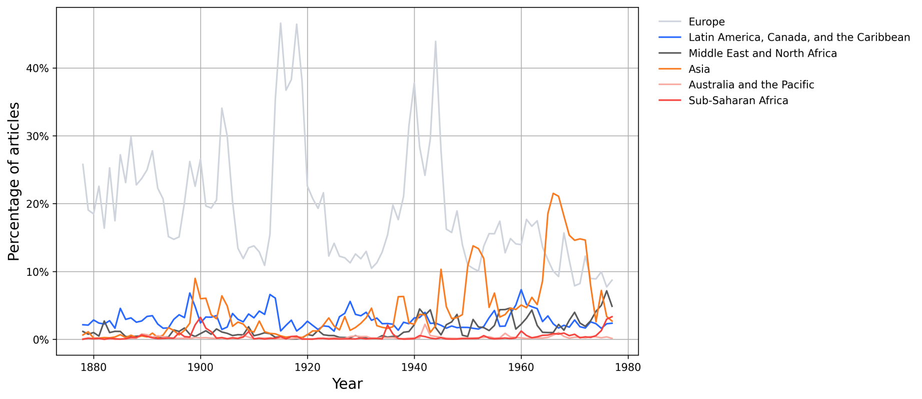
Geographic Coverage

- Washington DC: 27% of content
- New York: 5%



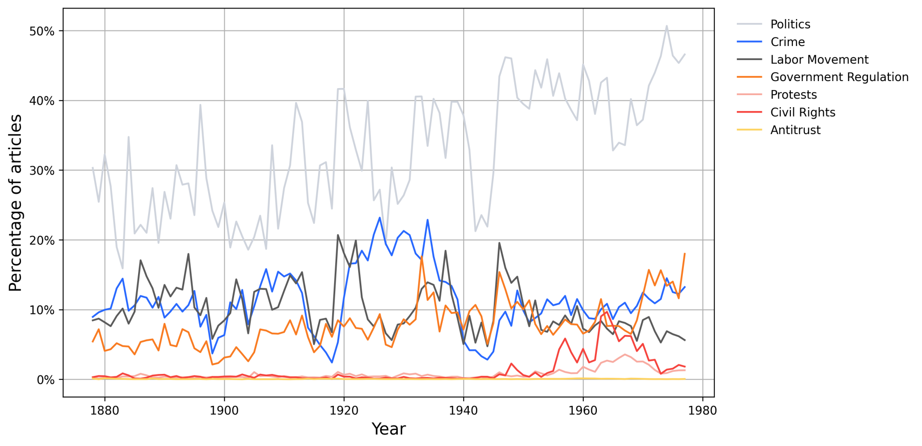
International Coverage

- 25.7% international datelines
- Peaks during World Wars
- Strong focus on Europe

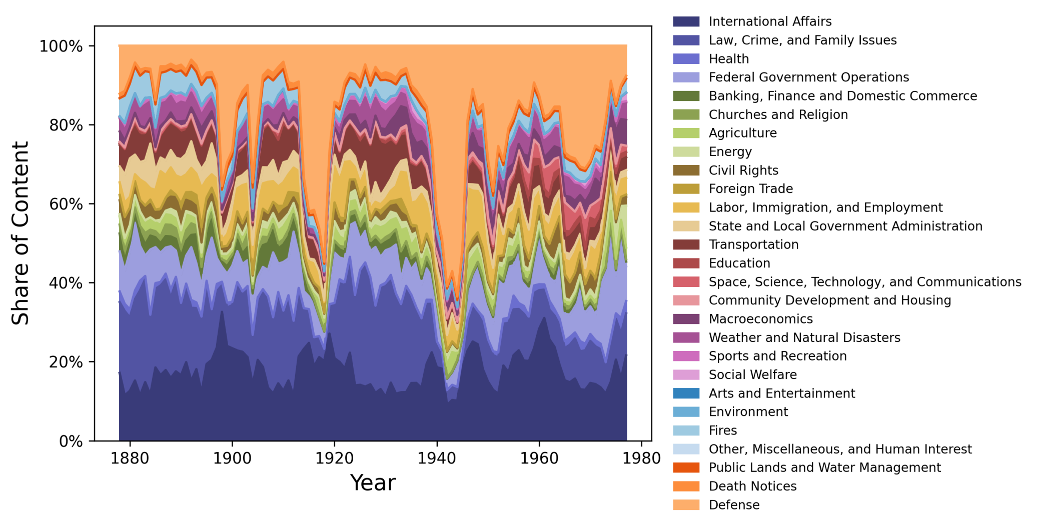


Topic Tagging

- Politics: 37% of articles
- Crime peaks during Prohibition
- Protests peak in 1960s

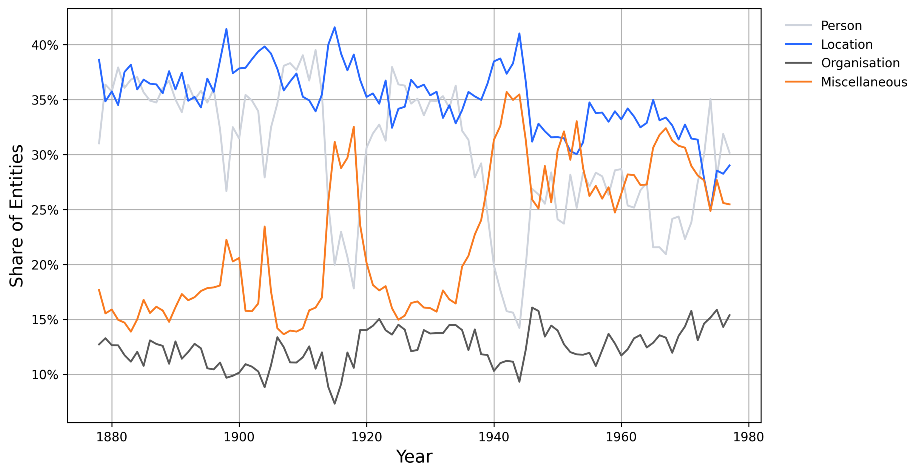


Comparative Agendas Topics



Named Entity Recognition

- 43.7M entity mentions
- Tracks major historical events eg. WWI and WWII



Entity Disambiguation

- 15.3M person mentions disambiguated to Wikidata
- 61,933 unique individuals
- Most mentioned:
  - Dwight D. Eisenhower (9,530 articles)
  - Richard Nixon
  - Harry S. Truman
  - Adolf Hitler
  - Nikita Khrushchev
- Only 4.6% of mentions are women
- Most mentioned woman: Golda Meir

Applications

- Language Model Training
  - Additional historical information
  - Avoid look-ahead bias
  - Reduce copyright risk
- Retreival-augmented applications
- Research Applications
  - Computational linguistics
  - Social science
  - Digital humanities

Links and More Information

- Dataset available on Huggingface (dell-research-harvard/newswire)
- All models used also on huggingface
- For more information on methods and evaluation, see our paper.